De novo pathway-based biomarker identification

Alcaraz, Nicolas; List, Markus; Batra, Richa; Vandin, Fabio; Ditzel, Henrik; Baumbach, Jan

Published in:
Nucleic Acids Research

DOI:
10.1093/nar/gkx642

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Document license
CC BY-NC

Citation for pulished version (APA):
Alcaraz, N., List, M., Batra, R., Vandin, F., Ditzel, H. J., & Baumbach, J. (2017). De novo pathway-based biomarker identification. Nucleic Acids Research, 45(16), [e151]. DOI: 10.1093/nar/gkx642

Published online 20 July 2017

Nucleic Acids Research, 2017, Vol. 45, No. 16 ***e151***
doi: 10.1093/nar/gkx642

De novo pathway-based biomarker identification

Nicolas Alcaraz[1,2,3,*,†], Markus List[4], Richa Batra[5,6], Fabio Vandin[1,7], Henrik J. Ditzel[2,8] and Jan Baumbach[1,9]

[1]Department of Mathematics and Computer Science, University of Southern Denmark, 5230 Odense, Denmark, [2]Department of Cancer and Inflammation Research, Institute of Molecular Medicine, University of Southern Denmark, 5000 Odense, Denmark, [3]The Bioinformatics Centre, Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark, [4]Computational Biology and Applied Algorithms, Max Planck Institute for Informatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany, [5]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Munich, Germany, [6]Department of Dermatology and Allergy, Technical University of Munich, 80802 Munich, Germany, [7]Department of Information and Engineering, University of Padowa, 35122 Padowa, Italy, [8]Department of Oncology, Odense University Hospital, 5000 Odense, Denmark and [9]Computational Systems Biology Group, Max Planck Institute for Informatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany

Received March 14, 2017; Revised July 11, 2017; Editorial Decision July 13, 2017; Accepted July 13, 2017

ABSTRACT

Gene expression profiles have been extensively discussed as an aid to guide the therapy by predicting disease outcome for the patients suffering from complex diseases, such as cancer. However, prediction models built upon single-gene (SG) features show poor stability and performance on independent datasets. Attempts to mitigate these drawbacks have led to the development of network-based approaches that integrate pathway information to produce meta-gene (MG) features. Also, MG approaches have only dealt with the two-class problem of good versus poor outcome prediction. Stratifying patients based on their molecular subtypes can provide a detailed view of the disease and lead to more personalized therapies. We propose and discuss a novel MG approach based on *de novo* pathways, which for the first time have been used as features in a multi-class setting to predict cancer subtypes. Comprehensive evaluation in a large cohort of breast cancer samples from The Cancer Genome Atlas (TCGA) revealed that MGs are considerably more stable than SG models, while also providing valuable insight into the cancer hallmarks that drive them. In addition, when tested on an independent benchmark non-TCGA dataset, MG features consistently outperformed SG models. We provide an easy-to-use web service at http://pathclass.compbio.sdu.dk where users can upload their own gene expression datasets from breast cancer studies and obtain the subtype predictions from all the classifiers.

INTRODUCTION

High-throughput gene expression profiling from DNA microarrays in combination with machine learning techniques is a widespread approach to identify genes that can be used to stratify patients into groups with distinct clinical outcome. These so-called gene expression panels (GEPs) can then be used to guide clinicians in selecting the appropriate therapy in complex diseases, for example, lupus (1), Chron's disease (2) or Parkinson's (3). Commercialized GEPs have already been developed for outcome prediction in cancer: Mammaprint® (4), a set of 70 genes for low- or high-risk prediction in breast cancer, Decipher® (5), a panel of 22 RNA markers to predict risk of metastases in prostate cancer and OncoType DX® panels for tumor profiling in breast (6), prostate (7) and colon (8) cancer. Although GEPs have shown relative success in cancer prognosis, it was demonstrated that their predictive performance is not consistent across datasets (9). An insufficient number of samples, inherent noise in the experiments and the heterogeneity of cancer patients have been pointed out as main reasons for the lack of feature stability and prediction accuracy. Moreover, cross validation shows that even on a single dataset, the size and composition of the selected gene panels varies strongly (10). The large number of features, which are often correlated, pose a significant challenge in cancer subtyping when only comparably few samples are available ($p \gg n$ problem).

Complex diseases such as cancer have thus driven the need for more 'systems'-based approaches that elucidate the molecular mechanisms underlying the disorder rather than the effect of individual genes. Hence, recent efforts to predict outcome in cancer exploit the wealth of information about protein–protein interactions, gene regulations,

*To whom correspondence should be addressed. Tel: +45 65502309; Email: nalcaraz@binf.ku.dk
†These authors contributed equally to the paper as first authors.

metabolic reactions and other types of relationships between biomolecules available in public databases. In the context of patient outcome prediction, a popular approach is to aggregate groups of biologically related genes into gene sets with a summary activity score, often called meta-genes (11) (MGs). They are hence also called 'composite-features' (12), which are used in supervised learning methods in the expectation that they will provide improved prediction performance and higher feature stability compared with single-gene (SG) panels (i.e. sets of genes treated as *a priori* unrelated and independent). MGs can also be used to group highly correlated genes and reduce the overall feature number, thus alleviating the $p \gg n$ problem. In addition, MGs representing pathways can provide a higher level of interpretation and help biomedical researchers to identify novel biomarkers and drug targets.

To obtain a set of predictive MGs in cancer, some approaches (13,14) use predefined pathways or gene sets expertly curated and stored in public databases (e.g. Kyoto Encyclopedia of Genes and Genomes (KEGG) (15), Reactome (16)). However, most *a priori* defined pathways are not disease-specific and are thus only partially affected during the course of the disease in question. Despite continuous improvement over the years, interaction databases are still biased toward well-understood biological processes and few pathways are compiled and annotated for rare or specific (sub-)types of diseases (17). Other methods (18,19) extract a new list of case-specific pathways by searching for connected subnetworks in a large interaction network. The top ranking subnetworks are then selected and used as MG features for the classification procedure. We previously termed this pathway extraction step as *de novo* pathway enrichment, (20), although other terms exist as well such as functional module detection or connected subnetwork extraction. Various *de novo* pathway enrichment methods have been proposed (e.g. (21–26)) differing in subnetwork scoring function, optimization criteria or search method. Nevertheless, to the best of our knowledge, popular and well-established methods have never been used in the context of pathway-based prediction. We believe this is due to limiting factors such as high runtime, low robustness (27), unintuitive parameters or lack of a publicly available software package or web service (23,28) that can be easily integrated into the classification pipeline. For these reasons, it is understandable that most MG classifiers extract subnetworks using simple but fast search heuristic methods (19,29) that may be prone to overlook relevant genes or interactions.

Even though previous studies employing MG classifiers report an improvement over SG classifiers, other evaluations (30–33) have challenged such claims. In more rigorous and exhaustive simulations performed on larger datasets, Staiger *et al.* (32) demonstrated that MGs do not outperform SGs in prediction performance or stability over a range of different networks, gene sets or classifiers. In a more recent study (34), Allahyar *et al*. identified shortcomings that could lead to loss of predictive power of MG classifiers evaluated by Staiger *et al.* In particular, they argued that using simple averaging operators to produce a MG score can lead to loss of predictive power, while performing feature extraction and feature ranking for selection in separate steps can produce unstable features. Allahyar *et al*. show that by introducing certain improvements to the MG classifiers that mitigate these problems, their performance increases and in some cases achieve equal or better results than SG classifiers. Allahyar *et al*. further introduce FERAL, a method that produces multiple MGs from the same gene sets by using multiple aggregation operators. FERAL then implements a sparse group lasso approach that simultaneously selects the features and integrates them into the prediction model, thus avoiding the separate feature selection and ranking procedure. Although the authors report higher prediction accuracy with FERAL compared with other known MG classifiers as well as SGs, improvement over SGs is modest and it becomes evident that cancer outcome prediction based on gene expression might not have the potential to improve significantly regardless of the features, datasets or techniques used to build the classifiers (34).

Cancers are often characterized by the occurrence of subtypes with distinct clinical outcome. One example is found in breast cancer, in which clinically relevant subtypes have been defined that show significant differences in terms of their incidence, risk factors, prognosis and treatment sensitivity (35). In clinical practice, breast cancer is classified based on the estrogen and progesterone receptor status, as well as the expression of the Her2 gene. Nevertheless, histopathological classification of the tumor samples is prone to human error given that it must be performed by a skilled pathologist (36). In an effort to more reliably assign subtypes to cancer patients, the PAM50 (37) GEP was developed, a set of 50 genes stratifying patients into five 'intrinsic' molecular subtypes: Basal, Her2, LumA, LumB and Normal-like. Although efforts have been made to obtain smaller gene sets based on gene expression (38), the PAM50 gene panel currently remains the quasi-gold standard for breast cancer subtype classification and has been used, for instance, by The Cancer Genome Atlas (18) (TCGA).

In this article, we focus on a pathway-based prediction of cancer subtypes, as opposed to previous MG classifiers, which have only addressed the two class problem of good versus bad outcome prediction (based on time of death, recurrence or metastasis) in cancer patients. We propose a novel pipeline that extracts subtype-specific pathways suitable as features for classification models that can predict the subtype labels of patients. We measure performance, stability and functional enrichment of the most frequently selected features in a repeated cross-validation setting tested on a large cohort of breast cancer patients taken from TCGA (18). Subsequently, we show that MGs based on both *de novo* and *a priori* pathways are significantly more stable than SGs. Finally, we validated our MG models (learned from TCGA data) on a large and independent set of 12 breast cancer datasets from the Amsterdam Classification Evaluation Suite (32) (ACES) and show that *de novo* pathway features are able to predict the subtype labels with higher accuracy than SGs and pre-defined pathways, even on unseen data.

In our approach, feature extraction and MG score aggregation methods are independent of the statistical properties of the datasets and we demonstrate the applicability of the pipeline to other types of OMICs sets by using DNA methylation from the same TCGA cohort. Finally,

we show that frequently selected pathway-based features extracted from both, gene expression and DNA-methylation data, provide complementary enrichment for cancer hallmarks (39), which is not evident when focusing on the top selected genes in the SG models.

MATERIALS AND METHODS

Overview of the classification evaluation pipeline

We first provide an overview (Figure 1) and the motivation behind the design choices in each of the pipeline steps. Given a series of molecular profiles and large interaction network, the first phase consists of extracting subtype-specific pathways that can be used as MG features in the classifier (Figure 1B). For this task, we required a *de novo* pathway extraction method able to identify more than one subnetwork, given that multiple pathways are usually affected during cancer development and that one feature would be insufficient to distinguish between more than two classes. Also, to minimize feature instability, the *de novo* pathway enrichment method should be robust to noise in both the dataset as well as the network. Finally, run time can become an issue in large scale simulations, hence the tool should also be computationally efficient. We decided to use KeyPathwayMiner (KPM) (40), a *de novo* pathway enrichment tool that in addition to providing all the above features, has shown good performance compared with other tools (20). KPM is able to extract all maximal-connected subnetworks containing at most K genes not differentially expressed in at most L cases, and expects an indicator matrix as input in which '1' indicates differential expression or activity of a gene and '0' otherwise. Such a matrix can be computed with the most suitable statistical method for the given OMICs dataset type. The two parameters K and L serve to control the noise in the data by allowing for a certain number of outliers both in the measurements (L) and network (K). To extract pathways specific to a certain subtype with KPM, we produce a differential expression (methylation) indicator matrix for all pairs of the given subtype against all others and connect them via an 'AND' logical connector. In other words, we searched for maximal connected subnetworks containing genes that are differentially expressed against all other subtypes.

Once all sets of pathways have been extracted, the next step is to aggregate the single gene expression values into one summary score for each pathway (Figure 1C). We avoid traditional score aggregation operators (mean, median, etc.) which can lead to loss of information and instead employ single sample gene set enrichment analysis (41,42) (ssGSEA), a rank-based method for comparing the expression levels of genes in a gene set with all other genes in the expression profile for a single sample. As ssGSEA requires no phenotypic labels, one can compute the MG score for samples with unknown labels, a requirement for breast cancer subtyping in clinical settings. Furthermore, in contrast to other gene set enrichment methods (43,44), ssGSEA does not require information from the other samples to compute the final pathway score, which can be a source of overfitting in the training data. The final output of the ssGSEA method is a pathway versus samples matrix which was consequently used in the following statistical learning steps.

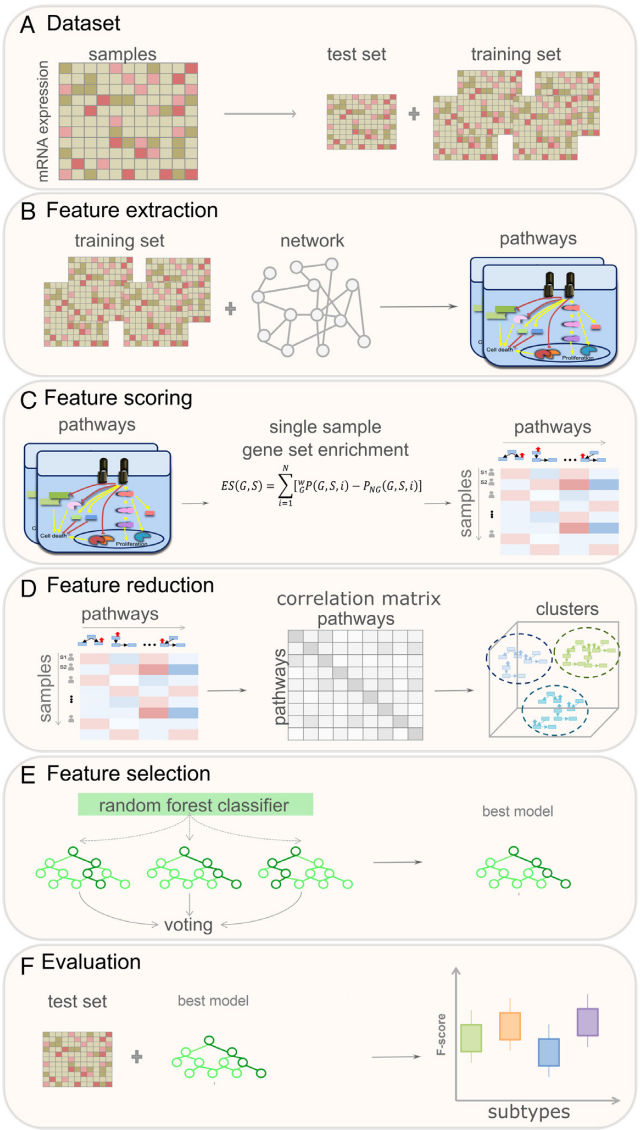

Figure 1. Workflow for *de novo* pathway-based classification of breast cancer subtypes. *Datasets* (**A**): dataset is split into 5-folds: four for training and one for testing. *Feature extraction* (**B**): subtype-specific pathways are extracted by projecting the training sets on the input network and running KeyPathwayMiner (KPM). *Feature scoring* (**C**): the extracted pathways are scored using single sample gene set enrichment analysis (ssGSEA) to produce a matrix of samples versus pathways. *Feature reduction* (**D**): pathways are clustered based on their Spearman's correlation coefficient across all training samples and the most representative feature for each cluster is selected for the next step. *Feature selection* (**E**): the best features are selected using random forests (RFs) with recursive feature elimination and a final RF model is built with the selected features. *Evaluation* (**F**): the final model is used to predict the subtype labels of the test set. The process is repeated for all fold splits and 10 repeats, recording the performance for each run.

Highly correlating features, a common problem in biomedical datasets, can reduce prediction performance or increase feature instability (45). To remove correlated features (Figure 1D), we computed their Spearman's correlation coefficient, clustered them using TransClust (46), which has shown good performance with biomedical datasets (47) and selected the most representative feature of each cluster. Afterward, we performed feature selection (Figure 1E)

with random forest (RF) models. Briefly, RF is a supervised learning algorithm that uses an ensemble of decision trees trained on bootstrapped samples of the data. The RF model has several advantages: good performance in multi-class scenarios, incorporates interactions between variables, requires little parameter tuning and per default returns measures of variable importance. In the case of breast cancer classification, RF models have been shown to be successful when applied on gene expression profiles (48,49). We employed the varSelRF (50) package to perform recursive feature elimination with RFs, which compared with the classical RF method, returns a small set of feature that retains high predictive accuracy. The final feature list was subsequently used to train an additional RF model that we used to predict the subtype labels of the independent test datasets.

In the case of *de novo* pathway models, all steps were performed inside a 10-times repeated 5-fold cross validation loop (Figure 1A). The data were split such that each fold kept the same proportion of samples for each subtype (stratified cross-validation). When known pathways or gene sets were available the features were fixed, hence the feature extraction phase was skipped and the scoring step (ssGSEA) only needed to be computed once before the cross validation loop. The SG models were constructed directly from the expression values, starting from the clustering step to remove correlating features (genes), while performing feature selection and building the final model within the same 5-fold CV scheme, with all steps using the same tools and parameter settings as with the MG models.

To avoid any unfair comparisons due to random fold splits, all models were trained and tested on exactly the same splits. In addition to the selected genes from the SG models, we built RF models using the 50 genes from the PAM50 gene set as features and evaluated them under the same repeated 5-fold cross validation scheme.

Gene expression and DNA-methylation datasets

Gene expression and DNA-methylation samples from breast cancer tumors were downloaded from the TCGA (18). All datasets were obtained in their processed level 3 form. Negligible batch effect was detected in the original analyses (18), hence no further batch correction was performed to avoid loss of biological signal. DNA-methylation probes mapping to the same gene were median centered to obtain a single value per gene. Afterward, the β-values were converted to M-values.

As additional gene expression source, the (32) ACES datasets were used as external validation set. They consist of a large cohort of breast cancer gene expression datasets from 12 different studies which already have been normalized, corrected for batch effects and filtered for duplicated samples. See Supplementary Table S1 for number of samples per PAM50 subtype. Batch correction was performed using the ComBat (51) function implemented in R package `sva` (52).

Gene sets and pathway sources

CPDB. The Consensus Path Database (53) is a collection of human pathway sources from 32 public databases comprising seven different types of associations: protein interactions, signaling reactions, metabolic reactions, gene regulations, genetic interactions, drug–target interactions and biochemical pathways. We downloaded all 4593 human gene sets (no interaction information is available), release 31, comprising the pathways with their Entrez gene identifiers. All pathways containing >300 genes representing very general biological processes were removed and 3906 pathways remained.

MsigDB. The Molecular Signature Database (42) stores different collections and sub-collections of biological relevant gene sets. We downloaded the C2 gene set, version 5.1, containing curated gene sets from pathways sources, biomedical literature and expert knowledge. Several of the gene sets are built based on microarray experiments of knockout studies. We downloaded all 4726 gene sets together with their Entrez gene identifiers.

KEGG cancer hallmark gene sets. To perform cancer hallmark enrichment, we collected a set of KEGG (15) pathways related to each cancer hallmark. See Supplementary Table S2 for details of all pathways.

Human interaction databases

HPRD. The Human Protein Reference Database (54) version 9 contains information of protein–protein interactions curated from literature. The protein–protein interaction network with Entrez gene identifiers contains a total of 9520 proteins and 39 227 interactions.

I2D. The I2D database (55), obtained in February 2015, contains protein–protein interactions curated from different interaction databases such BIND, HPRD and MINT as well as predicted interactions. We filtered out all predicted interactions and self-interactions. After converting Uniprot identifiers to Entrez gene identifiers a total of 15 379 proteins and 209 203 interactions remained.

HTRIdb. The Human Transcriptional Regulation Interaction Database (56), downloaded in September 2015, compiles experimentally validated transcription factor—target gene regulations in human. After removing all self-regulations and disregarding directional information, a total of 18 310 genes with their Entrez gene identifiers remained, with a total of 51 833 interactions.

HumanNet. The HumanNet (57), version 1.0, is a probabilistic functional gene network constructed with Bayesian integration of 21 types of OMICs data. Each interaction contains a log-likelihood score (LLS) measuring the probability that an interaction represents a functional relationship between two genes. We removed interactions with an LLS < 1.5, leaving a network containing 13 022 genes in Entrez gene identifiers and 123 052 edges with high confidence scores.

Pathway extraction with KeyPathwayMiner

Extraction of subtype-specific pathways was performed with KPM (40,58). For each subtype, an indicator matrix of

genes versus samples was produced as following: a *P*-value was computed for each gene and each sample in the given subtype by performing one sample Mann–Whitney U-test (two-sided) against the same gene and all samples in a different subtype. All *P*-values were corrected with Benjamini–Hochberg procedure and a '1' was placed in the corresponding entry in the indicator matrix if the adjusted *P*-value $<$ 0.05, all other entries were filled with zeroes. Once the set of indicator matrices of differential activity for the given subtype against each other subtype was obtained, a final subtype-specific matrix was produced by connecting all entries with an 'AND' logical operator. In other words, one in the matrix indicates that the gene is differentially active in that sample against all other subtypes. Afterward, KeyPathwayMiner was executed with the following parameters: Individual Node Exceptions (INEs) model, Greedy algorithm, Border Exception Node (BEN)-free option on, $K = 2$ and $L = 10\%$ of the cases. Hence, for each subtype and the given networks, all maximal connected subnetworks containing at most $K = 2$ genes not differentially active in at most $L = 10\%$ of the samples were extracted. The choice for K, L was based on 5-fold cross-validation runs performed over the TCGA dataset and the values for which the models showed the best average *F*-score over all networks were selected for the final pipeline evaluation.

Meta-gene scoring

In order to produce an activity matrix to score the gene sets representing pathways, we used the single sample gene set enrichment method (42) as implemented in the Gene Set Variation Analysis (GSVA) (43) R package with default parameters. For more information about the ssGSEA methods, see Supplementary Data.

Removal of correlating features

Features were clustered based on their Spearman's correlation coefficient using TransClust (46), with a threshold value of 0.9. This produced clusters of features where the average Spearman's correlation value of all pairs of features within each cluster was above 0.9. Finally, the feature with the highest average similarity within the cluster was taken as cluster representative, while all other features in the cluster were discarded from further workflow steps.

Feature selection and model building

To select a small set of predictive features we used the varSelRF R package, which performs feature selection with RFs using a recursive feature elimination approach. The feature selection procedure implemented in varSelRF starts by building an RF with all features and then iteratively proceeds to remove 20% of the least important features. This procedure is repeated until a model with only two features is left. The model with the lowest out-of-bag error (OOB) is reported as a final solution.

RESULTS

Performance comparison within TCGA gene expression datasets

To test our subtype classification pipeline and to compare the performance of different types of features, we downloaded a cohort of >500 gene expression breast cancer samples from TCGA (18). The sample information includes the PAM50 subtype classification gold standard, which were used as subtype labels for supervised learning. Due to their low abundance (eight samples), Normal-like subtype samples were removed. We measured prediction performance of all models with the *F*-score (the harmonic mean of precision and recall), which is well suited for unbalanced multi-class problems (59).

We observe that the PAM-50 gene features are the top performers in the gene expression datasets, achieving a median *F*-score of 0.87 (see Figure 2A). Surprisingly, the MGs from MsigDB gene sets perform almost equal to SG classifiers, both obtaining a median *F*-score of 0.85. This demonstrates that ssGSEA is able to capture the activity of the relevant genes in the MGs without loss of discriminative power. *De novo* pathway features extracted from HumanNET and HPRD show slightly lower performance than SG and MsigDB features but clearly outperform CPDB pathways. Features corresponding to HTRIdb and I2D are the lowest performing *de novo* pathway features and slightly outperform CPDB. When looking into the classification performance per subtype (see Figure 2C), we observe that *de novo* pathways and CPDB features suffer more in Her2 subtype prediction within the TCGA datasets, while performance in other subtypes is similar to SGs. The Her2 subtype is mostly misclassified into LumB and LumA subtypes, which can be due to overlap between Her2 clinical subtypes and Luminal mRNA subtypes (6,18) that affect similar pathways or regions in the network.

Validation on independent ACES datasets

To validate if the same findings would hold for independent datasets, we used the ACES benchmark datasets, a collection of over 1600 breast cancer samples from 12 different gene expression microarray studies available at the NCBI's Gene Expression Omnibus (60). The ACES datasets, which are also annotated with PAM50 labels, were already used for MG classification performance evaluation in studies by Staiger *et al.* and Allahyar *et al.* For each of the features, we trained a model on the full gene expression TCGA dataset, using exactly the same pipeline steps inside the cross-validation loop and predicted the PAM50 sample labels in the ACES datasets.

Despite the use of cross-validation, we observe a drop of performance for all models compared with the TCGA (Figure 2B and D). Most surprisingly *de novo* pathway features consistently outperform SG models. Another interesting finding is that MsigDB features outperform PAM50 genes on 9 out of 12 datasets (Supplementary Figure S1). This further demonstrates that grouping genes into MGs can reduce over-fitting and improve prediction performance by taking into account additional information provided by genes with lower discriminative power.

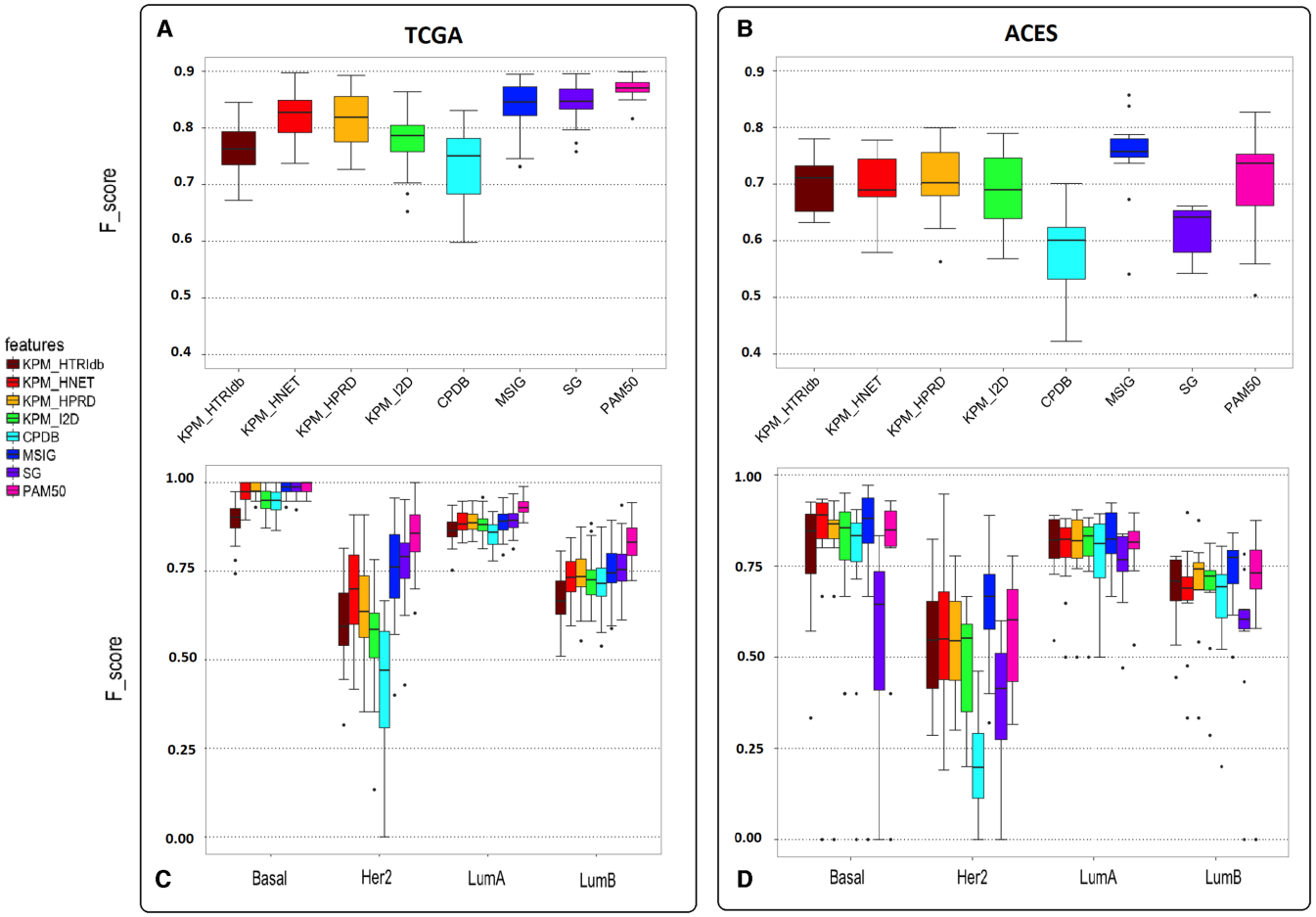

Figure 2. Prediction performance (*F*-score) for the different models within the TCGA cross-validation runs (**A** and **C**) and the final validation on the unseen ACES datasets (**B** and **D**) for gene expression. Top figures correspond to the overall performance and bottom figures to performance by class.

Note that we observed batch effects between the TCGA and ACES datasets—mainly in the Basal subtype (Supplementary Figure S15). After batch correction (Supplementary Figure S16), we observed an increase in performance for all feature types. Models trained with SG features were most affected (Supplementary Figures S17 and 18). This demonstrates that MGs are substantially more robust against batch effects.

While *de novo* extracted pathway features outperformed *a priori* defined pathways, we were curious why CPDB models did not perform as well as MsigDB ones. This can be explained by looking into the top features in their corresponding RF models after sorting by the mean decrease in accuracy. The top CPDB features (Supplementary Figure S8) only contain a few cancer-related pathways where some are quite general (e.g. Wikipathways_Integrated_Cancer_Pathway) and the rest unrelated to other cancer types. On the other hand, the top MsigDB features (Supplementary Figure S9) are dominated not by pathways but rather gene sets collected from breast cancer gene expression studies.

Feature and gene stability in gene expression

To evaluate feature stability for the SG and MG models from known pathways, we calculated the pairwise Jaccard Index between all pairs of folds in each run. We can see that in the gene expression datasets, CPDB and MsigDB features are considerably more stable than SGs (Figure 3). In the case of MGs from *de novo* pathways, stability at the feature level cannot be measured directly, since these are extracted anew during every run. Instead, we focused on the genes contained in the selected pathways and computed the stability at the gene level. We observe that, in this case, the Jaccard Index of genes in *de novo* pathways is similar to SGs, with the distribution slightly skewed to higher values for MGs than SGs (Figure 4A). However, when looking into the absolute selection frequencies of genes, we observe that a considerable higher number of genes were selected across all runs and folds in the *de novo* pathway features while only two genes were always selected in the SG models (Figure 4B).

DNA-methylation results

To gain insight into other molecular mechanisms that drive breast cancer, we executed the pipeline on the TCGA breast

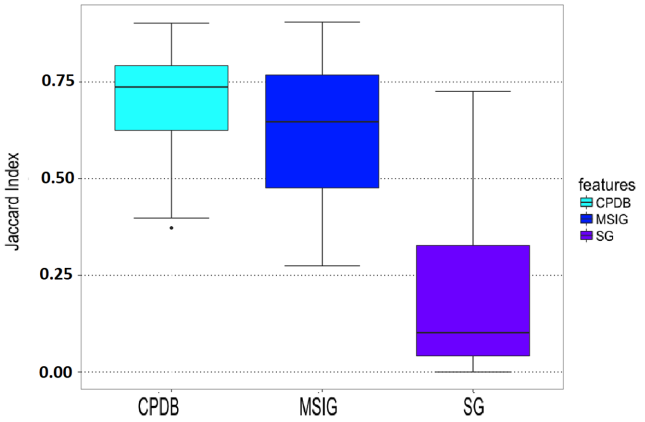

Figure 3. Stability of MG features from *a priori* pathways inside the cross-validation evaluation scheme for the TCGA gene expression cohort. The Jaccard Index was computed for the selected features of each pair of folds.

cancer DNA-methylation datasets. Given that the PAM50 labels are defined based on gene expression data, we observe an expected drop in performance for all models compared with the mRNA datasets. In particular, the 50 genes from PAM50 and the features extracted from the HTRIdb network are outperformed by all other models (Supplementary Figures S3a and b). To see if the performance of all the features was not just due to chance, we performed the same cross-validation pipeline with randomly permuted sample labels before each fold split (Supplementary Figure S5b). We observe that the performance drops for all features, demonstrating that DNA-methylation data also hold information on intrinsic subtype labels, but albeit to a lesser extent than gene expression data.

Cancer hallmark enrichment

If genes in *de novo* pathway markers are frequently selected, we expect them to be functionally enriched with cancer-related categories. To assess this, we compiled sets of genes from KEGG pathways related to each of the cancer hallmarks (see 'Materials and Methods' section and Supplementary Data). For all *de novo* pathway SG models, we selected the top 100 most frequently selected genes. Ties were broken by the average mean decrease in accuracy provided by the final RF models. Afterward, we performed a Fisher's exact test (same procedure as in (32)) for significant enrichment between the top genes and in each of the gene sets related to the cancer hallmarks.

Frequently selected genes in SG models were only significantly enriched in the 'genome instability and mutation' hallmark, whereas genes for all *de novo* pathway features had a higher enrichment score in that same category (Figure 5A). *De novo* pathway feature were additionally enriched in 'deregulating cellular energetics' (HTRIdb, I2D), 'resisting cell death' (HTRIdb, HNET and I2D) and 'sustaining proliferative signaling' (HTRDIdb, HNET, HPRD and I2D) hallmarks.

The same enrichment test was performed with the top 100 most frequently selected genes in the DNA-methylation datasets (Figure 5B). In this case we see an even more striking difference between SGs and MGs from *de novo* pathways. While SGs are not significantly enriched in any hallmark, MGs were enriched in four–seven hallmarks, depending on the MG model. Comparing results between expression and methylation, reveal that hallmark enrichment in MGs is complementary, i.e. hallmarks highly enriched in gene expression (e.g. 'resisting cell death' hallmark) are not found enriched in DNA methylation and *vice-versa* (e.g. 'tumor promoting inflammation' hallmark).

DISCUSSION AND CONCLUSION

Here, we present a novel approach for *de novo* pathway-based classification of breast cancer patients. We created the first publicly available online platform to provide multi-class breast cancer subtyping to the community.

We extracted subtype-specific *de novo* pathways from different human molecular biological networks, summarized their activity on a per-sample basis using ssGSEA and used them as features in an RF-based classification scheme. We compared prediction performance, stability and functional enrichment to other MG features from *a priori* defined pathways as well as to SG models in a repeated cross-validation setting applied to a large cohort of breast cancer samples from TCGA.

Performance and robustness

Our results show that SGs outperformed all MG models except for MsigDB gene sets, which achieved an almost equal *F*-score. However, when validated on the independent ACES datasets, performance of SGs dramatically decreased to a larger extent than *de novo* pathway features, demonstrating that using MGs as features for prediction models can reduce over fitting to the training data compared with SGs. Furthermore, MGs attained higher (MsigDB) or equally good (*de novo* pathways) *F*-scores than the PAM50 genes in the majority of ACES datasets, implying that grouping relevant genes into MGs can increase the prediction accuracy by exploiting the information contained in the related genes. It is evident that the choice of pathway source can have a significant impact on the prediction performance, as was shown by the large difference between CPDB and MsigDB features, both in TCGA and ACES datasets. We believe this is due to a lack of breast cancer specific gene sets in CPDB. Hence, *de novo* pathway features, which achieve comparable performance irrespective of the chosen interaction database, are preferable over *a priori* pathways in cases where few relevant gene sets or pathways are available for the disease under study.

Furthermore, robustness analysis also demonstrated that MG features from *a priori* pathways are remarkably more stable than SGs. The genes from *de novo* pathway features are more often selected as top genes in cross-validation simulations than genes from SGs indicating a higher robustness and thus, a higher breast cancer subtype relatedness of the genes in *de novo* pathways as compared with SG genes. Also, the top 100 most frequently occurring genes in *de novo* pathway features were significantly more enriched in cancer hallmarks than the top 100 genes in SG models. When comparing gene expression and DNA methylation, we found that

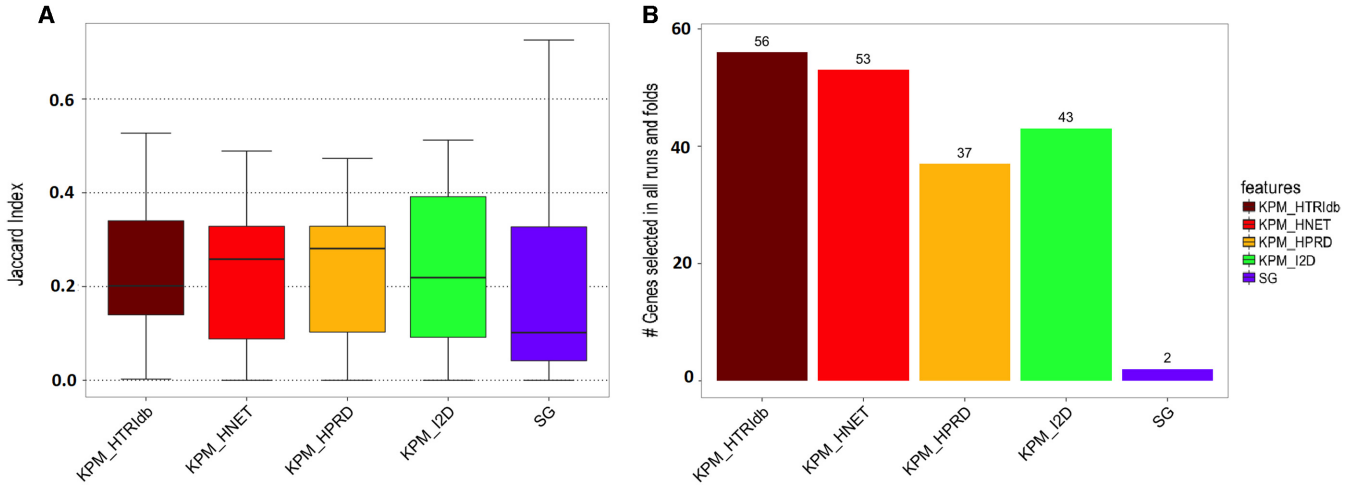

Figure 4. Stability of gene markers from *de novo* pathways inside the cross-validation evaluation scheme for the TCGA gene expression cohort. The Jaccard Index (**A**) was computed for the genes within the selected features for each pair of folds. In (**B**) the number of genes that were selected for all runs.

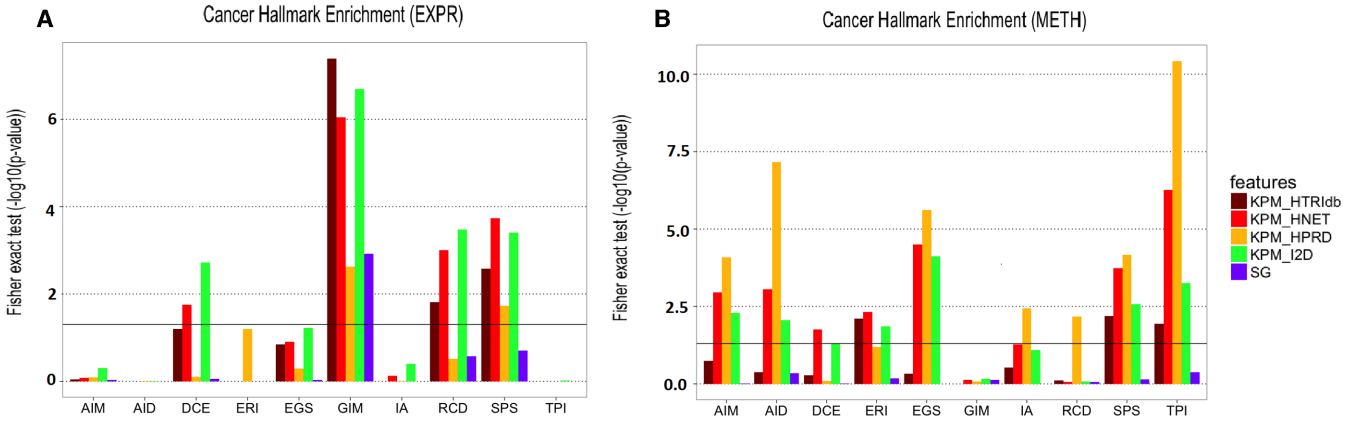

Figure 5. Cancer hallmark enrichment (Fisher's exact test) for the top 100 most frequently selected genes within the TCGA gene expression cross-validation loop in (**A**) gene expression and (**B**) DNA methylation. The cancer hallmarks are: activating invasion and metastasis (AIM), avoiding immune destruction (AID), deregulating cellular energetics (DCE), enabling replicative immortality (ERI), evading growth suppressors (EGS), genome instability and mutation (GIM), inducing angiogenesis (IA), resisting cell death (RCD), sustaining proliferative signaling (SPS) and tumor-promoting inflammation (TPI). Bold horizontal line corresponds to P-value = 0.05.

both types of data exhibited complementary cancer hallmark enrichment patterns, which are relevant information that were not evident in stable genes from SG models and highlights the advantage of integrating networks into cancer prediction models.

Multi-omics data analysis

By using DNA methylation from the same TCGA study, we also demonstrated the applicability of our method to other type of OMICs datasets. Performance of all models decreased, as expected, since PAM50 intrinsic subtype labels are based on gene expression (10). However, permutation tests based on label randomization showed that DNA-methylation levels of certain genes are also correlated, to a lower degree, to the PAM50 molecular subtypes. Nevertheless, SG models outperformed all other types of features, which can be due to a combination of factors such as a low number of subtype predictive genes found both in the networks and the methylation data. In addition, some interactions that constitute the networks can be based upon gene expression knockout studies (e.g. gene regulations) or other OMICs measurements uncorrelated to methylation patterns. Hence, differentially methylated genes tend to be sparse and distributed less centrally (61), which hampers the ability of *de novo* pathway enrichment methods to produce relevant subnetworks. Still, given these challenges, with F-scores of around 0.72 (Supplementary Figure S5) at least the MsigDB pathway-based models perform remarkably well and even similarly well compared with gene expression models when evaluated on the external ACES data (F-scores around 0.75, see Figure 2B).

Single-sample GSEA

We designed our pipeline by selecting tools that would avoid the pitfalls of previous network-based classification methods. To reduce possible loss of prediction performance, a crucial step is selecting the feature scoring method, where we consider ssGSEA to better utilize the information pro-

vided by a pathway compared with rather simple averaging operators. Moreover, ssGSEA does not tend to overfit to the dataset as it is the case for supervised methods that incorporate cross-sample information such as variance, correlation scores or phenotype labels. We increase feature stability by omitting a feature ranking step subsequent to feature scoring and instead perform robust feature selection with varSelRF to select the best features for the final model. In case of *de novo* pathway extraction, we use KPM, which efficiently extracts all pathways, irrespective of size, containing a high number of subtype-specific expressed genes. Finally, a common problem in network-based classifiers is highly correlated features. Our method uses TransClust to address this critical issue that was neglected in the past. This combination of tools allow us to perform network-based subtype classification providing more stable and interpretable features when compared with SGs, without compromising prediction performance.

Web service

We note that the modularity of our framework easily allows to replace individual tools at each step with others that could further increase prediction performance. However, we believe that the general observations we made would not change. Nevertheless, we plan to further investigate how different feature extraction, scoring and selection methods can affect performance and robustness.

Given that our pipeline applies a series of tools with one or more parameters, performing parameter optimization for all tools remains computationally prohibiting. Nevertheless, our main objective is to compare our MG classifiers versus the SG one in this multi-class scenario as unbiased as possible, hence we set the parameter values for most tools (TransClust, varSelRF and ssGSEA) to their recommended default values, which are based on their own evaluations on gene expression values. To the best of our knowledge, a systematic study on the parameter sensitivity of *de novo* pathway enrichment methods in classification scenarios is lacking. We thus decided to optimize KPM based on its performance on the TCGA dataset. Though we acknowledge that this may be a source of overfitting to the dataset, our final evaluation is based on an external dataset (the ACES dataset).

Network randomization

In addition, we ran our evaluation pipelines for randomized versions of the networks (Supplementary Figures S6 and 7). Similar to the findings in Staiger *et al.* and Allahayar *et al.*, we observe that models build from *de novo* pathways extracted from randomized networks do not decrease the performance significantly compared with the original networks. We hypothesize that the information in the biological network is provided by the global connectivity, such as the low average path length and scale-free degree distribution. Since these topological properties are not affected by standard random graph null models such as node label shuffling or degree preserving rewiring, the pathway extraction method is still able to find and group together relevant genes into pathways. To account for this in the future, we plan to extend the pathway extraction step and the pathway scoring method to also take into account the local pairwise connectivity of genes. This can provide further information in the form of confidence scores or correlation coefficients, potentially improving performance and robustness of the models.

Perspectives

An important next step is taking network-based disease sub-typing into the clinic. MG features have the advantage of being more robust against batch effects, which can not be realistically corrected for in a clinical setting, where a few or only one gene expression measurement may be available at a time. Another advantage of our framework is that once the classifier has been constructed and trained, it only requires the patient's gene expression measurements to predict its subtype. Similar to how GEPs have been defined and implemented in the clinic, we envision a future where clinicians would provide multi-gene expression measurements to software programs containing predictors that have been previously constructed from comprehensive datasets and high-quality biological networks. As proof-of-concept, we provide all datasets and classifiers at http://pathclass.compbio.sdu.dk including a small web service that allows for running all classifiers on user-uploaded gene expression and DNA-methylation datasets for online breast cancer subtyping.

Summary

In summary, the pipeline presented here demonstrated that MGs can be used for the purpose of subtype prediction in cancer. As features, MGs can provide higher accuracy, stability and biological relevance compared with SGs and can pave the way for better pathway-based classification techniques. In the future, we will extend our framework from single to multi-OMICs by integrating different sources of biological information.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

N.A. would like to acknowledge el Consejo Nacional de Ciencia y Tecnología (CONACyT) from Mexico for their financial support.

FUNDING

el Consejo Nacional de Ciencia y Tecnología (CONACyT) (to N.A.); ERC [676858-IMCIS to R.B.]; the Danish Cancer Society (to H.D.); National Experimental Therapy Partnership (NEXT) (in part); Innovation Fund Denmark ; VILLUM Young Investigator grant [13154 to J.B.]. Funding for open access charge: University of Southern Denmark.
Conflict of interest statement. None declared.

REFERENCES

1. Grammatikos,A.P., Kyttaris,V.C., Kis-Toth,K., Fitzgerald,L.M., Devlin,A., Finnell,M.D. and Tsokos,G.C. (2014) A T cell gene

expression panel for the diagnosis and monitoring of disease activity in patients with systemic lupus erythematosus. *Clin. Immunol.*, **150**, 192–200.
2. Arijs,I., Quintens,R., Van Lommel,L., Van Steen,K., De Hertogh,G., Lemaire,K., Schraenen,A., Perrier,C., Van Assche,G., Vermeire,S. *et al.* (2010) Predictive value of epithelial gene expression profiles for response to infliximab in Crohn's disease. *Inflamm. Bowel Dis.*, **16**, 2090–2098.
3. Molochnikov,L., Rabey,J.M., Dobronevsky,E., Bonucelli,U., Ceravolo,R., Frosini,D., Grünblatt,E., Riederer,P., Jacob,C., Aharon-Peretz,J. *et al.* (2012) A molecular signature in blood identifies early Parkinson's disease. *Mol. Neurodegener.*, **7**, 26.
4. van 't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
5. Karnes,R.J., Bergstralh,E.J., Davicioni,E., Ghadessi,M., Buerki,C., Mitra,A.P., Crisan,A., Erho,N., Vergara,I.A., Lam,L.L. *et al.* (2013) Validation of a genomic classifier that predicts metastasis following radical prostatectomy in an at risk patient population. *J. Urol.*, **190**, 2047–2053.
6. Paik,S., Shak,S., Tang,G., Kim,C., Baker,J., Cronin,M., Baehner,F.L., Walker,M.G., Watson,D., Park,T. *et al.* (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.*, **351**, 2817–2826.
7. Knezevic,D., Goddard,A.D., Natraj,N., Cherbavaz,D.B., Clark-Langone,K.M., Snable,J., Watson,D., Falzarano,S.M., Magi-Galluzzi,C., Klein,E.A. *et al.* (2013) Analytical validation of the oncotype DX prostate cancer assay—a clinical RT-PCR assay optimized for prostate needle biopsies. *BMC Genomics*, **14**, 690.
8. You,Y.N., Rustin,R.B. and Sullivan,J.D. (2015) Oncotype DX(®) colon cancer assay for prediction of recurrence risk in patients with stage II and III colon cancer: a review of the evidence. *Surg. Oncol.*, **24**, 61–66.
9. van Vliet,M.H., Reyal,F., Horlings,H.M., van de Vijver,M.J., Reinders,M.J. and Wessels,L.F. (2008) Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. *BMC Genomics*, **9**, 375.
10. List,M., Hauschild,A.C., Tan,Q., Kruse,T.A., Mollenhauer,J., Baumbach,J. and Batra,R. (2014) Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *J. Integr. Bioinform.*, **11**, 236.
11. Nevins,J.R. (2011) Pathway-based classification of lung cancer: a strategy to guide therapeutic selection. *Proc. Am. Thorac. Soc.*, **8**, 180–182.
12. Hou,D. and Koyuturk,M. (2014) Comprehensive evaluation of composite gene features in cancer outcome prediction. *Cancer Inform.*, **13**(Suppl. 3), 93–104.
13. Kim,S., Kon,M. and DeLisi,C. (2012) Pathway-based classification of cancer subtypes. *Biol. Direct*, **7**, 21.
14. Lee,E., Chuang,H.Y., Kim,J.W., Ideker,T. and Lee,D. (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, **4**, e1000217.
15. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
16. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F., McKay,S. *et al.* (2016) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.
17. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
18. Cancer Genome Atlas N. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
19. Chuang,H.Y., Lee,E., Liu,Y.T., Lee,D. and Ideker,T. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.
20. Batra,R., Alcaraz,N., Gitzhofer,K., Pauling,J., Ditzel,H.J., Hellmuth,M., Baumbach,J. and List,M. (2017) On the performance of de novo pathway enrichment. *NPJ Syst. Biol. Appl.*, **3**, 6.
21. Beisser,D., Klau,G.W., Dandekar,T., Muller,T. and Dittrich,M.T. (2010) BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics*, **26**, 1129–1130.
22. Breitling,R., Amtmann,A. and Herzyk,P. (2004) Graph-based iterative group analysis enhances microarray interpretation. *BMC Bioinformatics*, **5**, 100.
23. Ideker,T., Ozier,O., Schwikowski,B. and Siegel,A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**(Suppl. 1), S233–S240.
24. Nacu,S., Critchley-Thorne,R., Lee,P. and Holmes,S. (2007) Gene expression network analysis and applications to immunology. *Bioinformatics*, **23**, 850–858.
25. Qiu,Y.Q., Zhang,S., Zhang,X.S. and Chen,L. (2010) Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics*, **11**, 26.
26. Ulitsky,I. and Shamir,R. (2008) Detecting pathways transcriptionally correlated with clinical parameters. *Comput. Syst. Bioinform. Conf.*, **7**, 249–258.
27. Beisser,D., Brunkhorst,S., Dandekar,T., Klau,G.W., Dittrich,M.T. and Müller,T. (2012) Robustness and accuracy of functional modules in integrated network analysis. *Bioinformatics*, **28**, 1887–1894.
28. Ulitsky,I. and Shamir,R. (2007) Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.*, **1**, 8.
29. Taylor,I.W., Linding,R., Warde-Farley,D., Liu,Y., Pesquita,C., Faria,D., Bull,S., Pawson,T., Morris,Q. and Wrana,J.L. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.*, **27**, 199–204.
30. Cun,Y. and Frohlich,H.F. (2012) Prognostic gene signatures for patient stratification in breast cancer: accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics*, **13**, 69.
31. Staiger,C., Cadot,S., Gyorffy,B., Wessels,L.F. and Klau,G.W. (2013) Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Front. Genet.*, **4**, 289.
32. Staiger,C., Cadot,S., Kooter,R., Dittrich,M., Muller,T., Klau,G.W. and Wessels,L.F. (2012) A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. *PLoS One*, **7**, e34796.
33. Voyle,N., Keohane,A., Newhouse,S., Lunnon,K., Johnston,C., Soininen,H., Kloszewska,I., Mecocci,P., Tsolaki,M., Vellas,B. *et al.* (2015) A pathway based classification method for analyzing gene expression for alzheimer's disease diagnosis. *J. Alzheimers Dis.*, **49**, 659–669.
34. Allahyar,A. and de Ridder,J. (2015) FERAL: network-based classifier with application to breast cancer outcome prediction. *Bioinformatics*, **31**, i311–i319.
35. Yersal,O. and Barutca,S. (2014) Biological subtypes of breast cancer: prognostic and therapeutic implications. *World J. Clin. Oncol.*, **5**, 412–424.
36. Eccles,S.A., Aboagye,E.O., Ali,S., Anderson,A.S., Armes,J., Berditchevski,F., Blaydes,J.P., Brennan,K., Brown,N.J., Bryant,H.E. *et al.* (2013) Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. *Breast Cancer Res.*, **15**, R92
37. Parker,J.S., Mullins,M., Cheang,M.C., Leung,S., Voduc,D., Vickery,T., Davies,S., Fauron,C., He,X., Hu,Z. *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.
38. Prat,A., Parker,J.S., Fan,C. and Perou,C.M. (2012) PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. *Breast Cancer Res. Treat*, **135**, 301–306.
39. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
40. Alcaraz,N., List,M., Dissing-Hansen,M., Rehmsmeier,M., Tan,Q., Mollenhauer,J., Ditzel,H.J. and Baumbach,J. (2016) Robust de novo pathway enrichment with KeyPathwayMiner 5. *F1000Res.*, **5**, 1531.
41. Barbie,D.A., Tamayo,P., Boehm,J.S., Kim,S.Y., Moody,S.E., Dunn,I.F., Schinzel,A.C., Sandy,P., Meylan,E., Scholl,C. *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108–112.
42. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a

knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
43. Hanzelmann,S., Castelo,R. and Guinney,J. (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.
44. Tomfohr,J., Lu,J. and Kepler,T.B. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.
45. Tolosi,L. and Lengauer,T. (2011) Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, **27**, 1986–1994.
46. Wittkop,T., Emig,D., Lange,S., Rahmann,S., Albrecht,M., Morris,J.H., Bocker,S., Stoye,J. and Baumbach,J. (2010) Partitioning biological data with transitivity clustering. *Nat. Methods*, **7**, 419–420.
47. Wiwie,C., Baumbach,J. and Röttger,R. (2015) Comparing the performance of biomedical clustering methods. *Nat. Methods*, **12**, 1033–1038.
48. Kursa,M.B. (2014) Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics*, **15**, 8.
49. Statnikov,A., Wang,L. and Aliferis,C.F. (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, **9**, 319.
50. Diaz-Uriarte,R. (2007) GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics*, **8**, 328.
51. Johnson,W.E., Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (England)*, **8**, 118–127.
52. Leek,J.T. and Storey,J.D. (2008) A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 18718–18723.
53. Kamburov,A., Stelzl,U., Lehrach,H. and Herwig,R. (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D793–D800.
54. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
55. Brown,K.R. and Jurisica,I. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.*, **8**, R95.
56. Bovolenta,L.A., Acencio,M.L. and Lemke,N. (2012) HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, **13**, 405.
57. Lee,I., Blom,U.M., Wang,P.I., Shim,J.E. and Marcotte,E.M. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.
58. Alcaraz,N., Friedrich,T., Kotzing,T., Krohmer,A., Muller,J., Pauling,J. and Baumbach,J. (2012) Efficient key pathway mining: combining networks and OMICS data. *Integr. Biol. (Camb)*, **4**, 756–764.
59. Yu,H., Hong,S., Yang,X., Ni,J., Dan,Y. and Qin,B. (2013) Recognition of multiple imbalanced cancer types based on DNA microarray data using ensemble classifiers. *Biomed. Res. Int.*, **2013**, 239628.
60. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
61. Li,Y., Xu,J., Chen,H., Zhao,Z., Li,S., Bai,J., Wu,A., Jiang,C., Wang,Y., Su,B. and Li,X. (2013) Characterizing genes with distinct methylation patterns in the context of protein-protein interaction network: application to human brain tissues. *PLoS One*, **8**, e65871.